Cognitive Systems Evaluation Addendum
to the Proposer Information Pamphlet (PIP)

The paragraphs below enhance BAA 02-21 Modification 6, explaining many items in more detail.

DESCRIPTION

The intent of the Cognitive Systems Evaluation (CSE) Focal Challenge is to solicit proposals for the creation of a new capability to evaluate cognitive systems. The objective is to provide independent evaluations of cognitive system prototypes and related research and development performed for DARPA/IPTO. In general, the purpose of these evaluations will be to provide the necessary empirical evidence that research progress is being made toward the specific objectives of each relevant IPTO program. In addition these evaluation efforts may be used to verify that the capabilities under development for each program meet real user needs and additional requirements such as security, privacy, and trust, as needed by each individual program.

IPTO Mission and Cognitive Systems Evaluation

DARPA's Information Processing Technology Office (IPTO) has the mission of developing *cognitive systems*: computer systems that can reason, learn from experience, be told what to do, explain what they are doing, reflect on their experience, and respond robustly to surprise. Especially challenging is the task of evaluating these systems and their component technology, while that technology is being developed. The challenge is to devise scientifically valid evaluations that both accurately measure technical progress and provide a useful focus for technology development. Over the years, DARPA has sponsored Text REtrieval Conference (TREC), Message Understanding for Comprehension (MUC), speech understanding evaluations, robotic competitions, High Performance Knowledge Bases (HPKB) and Rapid Knowledge Formation (RKF) knowledge-base evaluations, and many more. It is difficult to devise evaluations that drive the development of the desired technology without encouraging the development of "throw-away" techniques that score well on the evaluation but do not solve the general problem. This challenge is even more severe when the target is the development of cognitive systems that are likely to involve significant human interaction and whose primary characteristic is the ability to respond intelligently to novel situations (and not only to perform well on precisely defined tasks).

Within this general context, IPTO has specific requirements for evaluating its research and development programs. The Perceptive Assistant that Learns (PAL) program will be developing and evaluating personalized cognitive assistants for general office tasks. The Personal Knowledge Pad (K-Pad) program will be developing and evaluating a "to-do" list assistant.
There are plans for new programs in reasoning, learning, human-computer interaction, and others that would all require thoughtful evaluation.

Scope of the Cognitive Systems Evaluation (CSE) Effort

The offeror will plan, design, and administer evaluations as well as analyze and interpret the data collected during these evaluations. The evaluation offeror will assess all aspects of the cognitive systems under development, including reasoning, perception, learning, advice taking, explanation, adaptation to new situations. Together with the various IPTO development contractors, the evaluation offeror will develop appropriate specifications, evaluation criteria, metrics and evaluation plans. The evaluations shall remain consistent from year-to-year in order to track year-to-year progress. However, the offeror shall adapt the evaluations specifications as each research program evolves. The initial contract for this solicited effort will be to provide independent evaluation of for the PAL (Perceptive Assistant that Learns) Program, explained below. It is IPTO's intention to add additional contract tasks in the future for the evaluation of other IPTO programs, such as the Personal Knowledge Pad, but the initial contract award will be for the evaluation of the PAL program.

Deleted: those

Deleted: ,

Independence of Evaluation

The Cognitive Systems Evaluation offerors must be able to offer independent evaluation of the Information Processing Technology Office performers' technologies and performance. That will require that the organization that is selected have complete independence in its evaluation process, and remain separated from the program performers. Ideally, offerors should be from organizations that would not expect to be performing technical work for IPTO, which would be subject to these evaluations. However, if an offeror which is performing or expects to perform technical work for IPTO can establish the independence of the organization conducting evaluation task from the organization performing technical tasks for IPTO, such an offer may be considered on a case-by-case basis.

General Steps in the Annual Evaluation Process:

In general, for most IPTO programs, there will be a 4-step evaluation process that will occur yearly:

- Task 1: Design, Specify, and Plan the Annual Evaluation: The offeror will design, specify, and plan the annual evaluations to be conducted for each IPTO R&D program under its purview. Although the evaluation offeror shall remain independent from the IPTO research and development contractors, the evaluation offeror shall work closely with the developers to specify evaluations that accurately validate the developer's claims, challenges, technology and accomplishments. Specific evaluations shall be focused on the interim products of each project. Evaluations shall include but not be limited to user-oriented measures of effectiveness and usability, as well as system-oriented measures of capability and cognitive technology performance. The central aspects of each technology will be evaluated with problems that reflect real-world challenges and constraints. The specification will define appropriate scenarios and evaluation problems to be implemented. Appropriate evaluation resources and infrastructure will be defined including evaluation data, messages, knowledge bases and human subjects. All human testing will be conducted following appropriate human-use guidelines and be reviewed by an Industrial Review Board (IRB) as noted below. The specification will also contain an evaluation design addressing the analytical methods and the nature of the domain samples to be examined. Statistical models will be formulated and customized as appropriate. The offeror shall develop concrete evaluation criteria and measures, to include both objective and subjective measures where appropriate. The offeror shall work with the development contractor(s) to address how any automated data collection schemes will be accomplished. The evaluation specifications will be maintained under configuration control and the offeror will ensure traceability. When new cognitive systems programs are identified by IPTO, the offeror will propose to the DARPA Program Manager an evaluation strategy for each emerging program.

- Task 2: Develop Evaluation Materials and Detailed Plan: The offeror shall plan, schedule, and develop appropriate evaluation materials for each annual evaluation activity and milestone and provide a more detailed evaluation plan during this task. Such scheduling shall be accomplished in coordination with DARPA and the development contractors. The detailed plan and evaluation material development activities shall reflect the overall plan and specifications developed in Task 1 above. The planning and evaluation materials shall be documented and maintained by the offeror.

- Task 3: Administer Evaluations: The offeror shall prepare and execute the evaluation, as well as perform the preliminary data summarization and validation. Preparation may include, but is not limited to, obtaining and sequestering equipment, knowledge bases and databases, conducting pilot evaluations, and dry runs, as needed. Execution activities include providing human subjects, evaluation administrators, evaluation monitoring and data recording resources. Evaluations may be conducted at the development contractor’s facility, at the evaluator’s facility, or at an appropriate military or operational site, depending on the nature of the test being conducted, as specified by the evaluation plan produced under Task 1. In all three cases, the offeror will be responsible for the overall administration and execution of the evaluation and for validating that all data was collected properly and accurately.

- Task 4: Analyze and Interpret Results: The offeror will analyze and interpret the results of the evaluations to identify the degree to which each program and technology claim has been substantiated. Technical causes of success or lack of success shall be identified. The data shall be interpreted with whatever statistical models were developed in the evaluation specifications developed during Task 1. The interpretation of the evaluation shall substantiate the system’s ability to observe and model the user’s context, improve over time, respond intelligently to novel situations, and naturally interact with users.

Deliverables

A. Evaluation Design, Specification, and Plan: The CSE offeror will develop and provide a document specifying the overall evaluation strategy, experimental design, and detailed evaluation plan for each evaluation. Generally, this evaluation specification should be delivered 12 months prior to the evaluation. This evaluation specification might typically include the following elements:

- Evaluation Overview and Purpose
  - Program goals
  - Technical claims to be tested
- Evaluation Strategy and Test Description
  - A description of the tasks to be performed
  - A description of the data, subjects, procedures to be used
  - Schedules for domain information, data, sample test problems, pilot studies, etc.
- Experimental Design
  - Hypotheses to be tested
  - Experimental and control conditions
  - Independent and Dependent Variables
  - Quantitative and Qualitative measures
- Experimental Method
  - Subjects
  - Training
  - Data
  - Experimental procedures and protocol
- Data Collection and Analysis Plan
  - How will the data be collected
  - How will the data be analyzed
  - What criteria (e.g., Go/No Go thresholds) will be applied
  - How will the analyzed results be interpreted

B. Evaluation Materials and Detailed Plan: An updated and more detailed plan for evaluation, including a detailed evaluation material plan, will be provided no later than 6 months prior to the evaluation in question. The more detailed evaluation plan will consist of the detailed tasks and activities to be pursued in conducting the evaluations, including detailed requirements for space, materials, and personnel from the participating performing contractors in accordance with Tasks 1 through 4 above.

C. Analytical Results: Within 3 weeks of each evaluation event the CSE offeror will provide to the DARPA Program Manager and the Program Manager's Government Agents a limited quick look of the evaluations performed by each performing contractor, consisting of summarized results of the evaluation and observations of the processes of evaluation. A full analysis will be required within 2 months of completion of the evaluation event, and should be based on a more thorough analytical foundation than the quick look report. In addition to a detailed assessment of which technology areas did and did not meet their goals, a more refined set of analyses detailing issues encountered during the evaluation and mitigating or enhancing factors that may have affected the evaluation of any given technology area should be developed. A specific set of recommendations should be included in the full analysis, to include, but not be limited to specific recommendations regarding the specific technology paths in subject program; recommendations regarding future evaluation design, planning and conduct; and recommendations regarding the evaluation process as a whole.

D. Documentation: In addition to the documentation defined in the main body of the PIP, the offeror will also deliver an evaluation specification, evaluation plan, evaluation materials and documented evaluation results as noted in Deliverables A through C above. The offeror shall also provide briefings to DARPA and at periodic DARPA Principal Investigator Meetings which will include prior year evaluation results, analysis, and up to date current year evaluation planning.

Initial Effort of the CSE, The PAL Program

The initial task of this solicited effort will be to provide independent evaluation of the PAL program. PAL (Perceptive Assistant that Learns) is a long-term DARPA R&D program that will greatly advance the field of cognitive computing and generate new ways for computers to support human activity. The PAL program is focused on developing an enduring personalized cognitive assistant which will make major advances in each of the following primary technology focus areas: Learning, Representation and Reasoning, Communications and Interaction, and Computational Perception. Underlying mathematical and scientific foundations supporting cognitive information processing will also be explored, resulting in theories that provide guidance in representation and control of reasoning systems.Greater description of the PAL program can be found on:

Deleted: for

http://www.darpa.mil/ipto/solicitations/open/02-21_Mod6.htm

SRI International and Carnegie Mellon University (CMU) are being funded by DARPA to design and develop prototype cognitive assistants that meet the requirements of decision-makers, including utility, security, privacy, and trust.

Carnegie Mellon University’s RADAR (Reflective Agents with Distributed Adaptive Reasoning) will help busy managers to cope with time-consuming tasks such as organizing their E-mail, planning meetings, allocating scarce resources such as office space, maintaining a web site, and writing quarterly reports. Like any good assistant, RADAR must learn by interacting with its human master and by accepting explicit advice and instruction. The RADAR project draws on Carnegie Mellon's expertise in artificial intelligence, machine learning, natural-language understanding, and human-computer interaction.

The SRI project is named CALO. SRI’s CALO, which will learn by working with, observing, and being advised by its users, will handle a broad range of interrelated decision-making tasks that have in the past been resistant to automation. It will have the capability to engage in and carry out routine tasks, and to assist when the unexpected happens. To accomplish all this, the CALO research team employs techniques from many fields: machine learning, human-computer interaction, natural-language processing, optimization, knowledge representation, flexible planning, and behavioral studies of real human managers and will be organized in terms of Technology Focus Centers (TFCs). SRI's CALO and CMU's RADAR efforts are further described in the web site:

http://www.darpa.mil/ipto/solicitations/open/02-21_Mod6.htm

Additionally, SPAWAR Systems Center (SSC), San Diego is performing both technical and contractual oversight of the PAL project, and will take an active role in the evaluation function. The nature of SSC's technical management role in the evaluation process has not been defined in detail, and will be refined upon selection of the Cognitive Systems Evaluation offeror.

The milestones for PAL evaluation are as follows:

- Initial Evaluations (Year 1): In the first year, 2004, of the PAL program, each performing contractor team will begin the development of the individual technologies, the computing infrastructure, and the integration and evaluation environment needed for PAL. The SRI and CMU teams will conduct evaluations of their progress which will only be observed and independently analyzed by the independent evaluator.

- Annual evaluation (Years 2-5): Beginning in the second year of the program, each PAL team will undergo an annual evaluation of their system against a problem that stresses the system against the full complexity of the real world. Both SRI and CMU have proposed evaluations, which have been reviewed and tentatively approved by DARPA. Each team will evaluate their PAL systems against separate evaluation problems. This same problem will be administered each year in order to measure PAL's progress as an enduring, personalized, cognitive assistant. It is expected that in the early years (years 2 and 3) PAL will struggle and demonstrate only minimal progress, but that in the later years (years 4 and 5), as the full suite of new technology is developed, PAL will demonstrate mastery over the problem. The evaluation problems and metrics for the SRI Internatrional and CMU teams are further described in the web site: http://www.darpa.mil/ipto/solicitations/open/02-21_Mod6.htm

Tasks specifically for the PAL program. These task descriptions modify or supplement those provided above and do not replace them:

- Task 1: Design, Specify, and Plan the Annual Evaluation: SRI and CMU have already established an evaluation framework and are contracted to administer their own respective evaluations. The independent evaluator will specify detailed evaluation conditions, evaluation questions and related material within the respective SRI and CMU frameworks. The independent evaluator will develop a refined Evaluation Design, Specification, and Plan jointly with CMU and SRI.

- Task 2: Develop Evaluation Materials and Detailed Planning: No modification to Task 2 above.

- Task 3: Administer Evaluations: The PAL performing contractors will prepare and execute the Year 1 evaluation, and have continuing execution responsibilities in their contracts. The CSE offeror will independently monitor and validate the evaluations.

- Task 4: Analyze and Interpret Results: No modification to Task 4 above.

Deliverables for the PAL Program These deliverables modify or supplement those deliverables detailed above and do not replace them altogether

A. Evaluation Design, Specification and Plan: An evaluation specification, design, and plan, in accordance with the guidance in task 1 above will be delivered each year, starting on or about the anniversary of the PAL contracts in May of 2004 (assuming a 1 February 2004 start date for the CSE contract). The May 2004 evaluation specification and design delivery will define the Year 2 evaluations, scheduled to take place in May 05.

B. Evaluation Materials and Detailed Plan: A detailed evaluation plan, including a detailed evaluation material plan, will be developed and provided for each Year 2 through Year 5 evaluation and will be provided no later than 6 months prior to the evaluation in question. For example, Year 2 detail evaluation and evaluation material plan will be provided in November of 2004 for a May 2005 evaluation event.

C. Analytical Results. Within 3 weeks of the Year 1 evaluation event, which the CSE offeror will observe but not plan or administer, the CSE offeror will provide a limited quick look of the evaluations performed by each PAL performing contractor, consisting of observations of the processes of evaluation used by each PAL performing contractor. A full analysis will be required within 2 months of completion of the evaluation event.

D. Documentation: No additional or modifications of Deliverables above.

Milestones of the Cognitive Systems Evaluation Offeror for the PAL Program

Proposals submitted in response to the CSE Focal Challenge must specifically address milestones at the annual PAL anniversary points (May of 2004 through 2008). The establishment of milestones, while at the discretion of the proposer, should clearly provide demonstrable evaluation of the capabilities cumulatively achieved by the team at the milestone described and meet the schedule and requirements described above. Proposals must discuss the use of phase or option years, to include fully defining the meaning of phase or option years. Base and phase or option years must be fully priced, to the extent possible. Deliverables, milestones, and demonstrations must be included and clearly defined with links to the Statement of Work. Proposers may propose a multiorganizational but integrated team. Submitted proposals must clearly discuss the offeror's planned approach to the sharing of information and responsibilities with each PAL performer and other members of its own team. Each team will evaluate their PAL systems against separate evaluation problems. The Cognitive Systems Evaluation offeror, as an independent third party, will design, develop, and audit the evaluation procedures for both teams.

Period of Performance, Award Intentions, and Anticipated Funding Level.

The period of performance of this contract to support the PAL program is anticipated to be approximately 57 months, commencing with award early in FY 04, and completing with full analysis of the Year 5 evaluations. The periods of performance, detailed delivery schedules, and analytical products for evaluation of additional programs will be separately negotiated annually, contingent on the appropriation of funds.

Intentions are to award a single contract for the Cognitive Systems Evaluation function for the PAL evaluation tenure, although multiple awards might be considered for added IPTO program evaluations. Teaming with multiple entities, particularly if specific areas of expertise are enhanced through the teaming process is encouraged, though not mandatory. The funding level for this project will be phased to meet the schedule of evaluation. The first phase will be from contract award (estimated to be 01 Feb 04) through June of 04 and is estimated to be funded at approximately $500k, Thereafter, each phase year is estimated to be funded not more than $1 million per year, and will be phased to coincide with annual evaluations, including the quick look analysis following the examinations.

Qualifications of Personnel

The intent of the CSE Focal challenge is to foster a multi-disciplinary broad approach to cognitive systems evaluation that significantly advances the state of the art both at the system level and in core technology. It is essential that technologies, components, and techniques developed in the course of this research be general enough to lend themselves to other applications (portability to other cognitive applications)—the goal (as is the goal of the entire BAA) is to create powerful and reusable cognitive information processing technologies and techniques rather than simply to develop a single use evaluation capability for the PAL program or other specific IPTO programs. However, successful evaluation of the cognitive information processing technologies and components specifically noted in this BAA is paramount. As previously noted above, a key challenge faced in evaluation of such cognitive systems is the evaluation of systems that are highly interactive with humans, exhibiting some human-like behavior (e.g., learning, perception, and cognition). IPTO systems emulate human abilities, live in human environments with human users, and rely on human knowledge. The required expertise for evaluation of IPTO systems is broad and multi-disciplinary, including artificial intelligence (AI) (particularly natural language, learning, and knowledge technologies); cognitive science, psychology, experiment design; statistics and data analysis; modeling; simulations and problem generators; and ability to instrument the work environment, including software packages.

Other Requirements.

The evaluation factors listed in the basic PIP will remain in effect for this modification. The proposal must specifically address the proposer and team members' Intellectual Property rights, Government rights to Intellectual Property resulting from the research, and the methods to be used to meaningfully exchange possibly proprietary information among the awardees. The proposer must specifically address the use of human subjects, if any, at any point during the life of the project. If human subjects are not to be used, a statement of that fact must be included. If human subjects are to be used at any time in the project, the proposer must meet all requirements of Title 45, Code of Federal Regulations (CFR), Part 46 Protection of Human Subjects, effective 13 December 2001, Department of Health and Human Services. The Common Rule (Federal Policy) for the Protection of Human Subjects (56 FR 28003) may also be found at 32 CFR Part 219, Department of Defense. If human subjects are to be used at any time during the project, the proposal must include Industrial Review Board (IRB) approval for the use of human subjects, or a plan for obtaining IRB approval for the use of human subjects prior to the use of the human subjects. Although proposals identified for funding under this Focal Challenge may result in a contract, grant, cooperative agreement, or other transaction depending upon the nature of the work proposed, the required degree of interaction between parties, and other factors, the Government anticipates awarding only contracts in order to maintain the desired level of control over this research.

GENERAL INFORMATION: The Government intends to use non-Government personnel to assist as special resources to assist with the logistics of administering the proposal evaluation and providing selected technical assistance related to proposal evaluation. Support personnel are restricted by their contracts from disclosing proposal information for any purpose. Contractor personnel are required to sign Organizational Conflict of Interest/Non-Disclosure Agreements. By submission of its proposal, each offeror agrees that proposal information may be disclosed to those selected contractors for the limited purpose stated above. Any information not intended for limited release to support contractors must be clearly marked and segregated from other submitted proposal material.

Deleted:

For those proposals submitted directly in response to this modification to BAA 02-21, Focal Challenge: Cognitive Systems Evaluation, a submission date of 12:00 NOON (ET), 12 December 2003 is established for consideration during the initial evaluation phase. Proposal cover sheets, as described in the basic BAA and Proposer's Information Pamphlet, should indicate the Cognitive Systems Evaluation Focal Challenge as the BAA technology area for which the proposal is submitted. Proposals should be submitted to the addresses contained in the basic BAA. The 12 December 2003 submission date applies only to proposals submitted in response to CSE Focal Challenge. Due to the extremely short submission time, no abstracts responding to this Focal Challenge are required or will be accepted. All other submission requirements, formatting

Deleted: 08

Deleted: 08

requirements, and reporting requirements remain the same as found in the initial solicitation describing the BAA. Offerors are still encouraged to submit other proposals addressing the issues in the basic BAA and to periodically check the Federal Business Opportunities and the IPTO Solicitation web page for possible new Focal Challenges and other modifications. As previously stated, the full proposal (original and designated number of hard and electronic copies) in response to this Focal Challenge must be submitted in time to reach DARPA by 12:00 noon (ET) Friday, 12 December 2003, in order to be considered during the initial evaluation phase. However, BAA 02-21 will remain open until 12:00 noon (ET) Monday, June 7, 2004. Thus, proposals for CSE Focal Challenge may be submitted at any time from issuance of this BAA amendment through Monday, June 7, 2004. While the proposals submitted after the Friday, 12 December 2003, deadline will be evaluated by the Government, proposers should keep in mind that the likelihood of funding such proposals is significantly less than for those proposals submitted in connection with the initial evaluation and award schedule for this Focal Challenge.

Deleted: Mon

Deleted: 08

Deleted: Mon

Deleted: 08